**Tor Thogersen**

**DA 410 -- Multivariate Analysis -- Winter 2018**

**Project 1**

**Part 1:** Download airpoll.txt In this problem, we will only focus on the first 16 observations (cities). Read the data into R (as a data frame) and name the data as airpol.full

> my.datafile <- tempfile()

> cat(file=my.datafile, "

+ akronOH 36 11.4 3243 8.8 15 59 921.9

+ albanyNY 35 11.0 4281 3.5 10 39 997.9

+ allenPA 44 9.8 4260 0.8 6 33 962.4

+ atlantGA 47 11.1 3125 27.1 8 24 982.3

+ baltimMD 43 9.6 6441 24.4 38 206 1071.0

+ birmhmAL 53 10.2 3325 38.5 32 72 1030.0

+ bostonMA 43 12.1 4679 3.5 32 62 934.7

+ bridgeCT 45 10.6 2140 5.3 4 4 899.5

+ bufaloNY 36 10.5 6582 8.1 12 37 1002.0

+ cantonOH 36 10.7 4213 6.7 7 20 912.3

+ chatagTN 52 9.6 2302 22.2 8 27 1018.0

+ chicagIL 33 10.9 6122 16.3 63 278 1025.0

+ cinnciOH 40 10.2 4101 13.0 26 146 970.5

+ clevelOH 35 11.1 3042 14.7 21 64 986.0

+ colombOH 37 11.9 4259 13.1 9 15 958.8

+ dallasTX 35 11.8 1441 14.8 1 1 860.1

+ daytonOH 36 11.4 4029 12.4 4 16 936.2

+ denverCO 15 12.2 4824 4.7 8 28 871.8

+ detrotMI 31 10.8 4834 15.8 35 124 959.2

+ flintMI 30 10.8 3694 13.1 4 11 941.2

+ ftwortTX 31 11.4 1844 11.5 1 1 891.7

+ grndraMI 31 10.9 3226 5.1 3 10 871.3

+ grnborNC 42 10.4 2269 22.7 3 5 971.1

+ hartfdCT 43 11.5 2909 7.2 3 10 887.5

+ houstnTX 46 11.4 2647 21.0 5 1 952.5

+ indianIN 39 11.4 4412 15.6 7 33 968.7

+ kansasMO 35 12.0 3262 12.6 4 4 919.7

+ lancasPA 43 9.5 3214 2.9 7 32 844.1

+ losangCA 11 12.1 4700 7.8 319 130 861.8

+ louisvKY 30 9.9 4474 13.1 37 193 989.3

+ memphsTN 50 10.4 3497 36.7 18 34 1006.0

+ miamiFL 60 11.5 4657 13.5 1 1 861.4

+ milwauWI 30 11.1 2934 5.8 23 125 929.2

+ minnplMN 25 12.1 2095 2.0 11 26 857.6

+ nashvlTN 45 10.1 2082 21.0 14 78 961.0

+ newhvnCT 46 11.3 3327 8.8 3 8 923.2

+ neworlLA 54 9.7 3172 31.4 17 1 1113.0

+ newyrkNY 42 10.7 7462 11.3 26 108 994.6

+ philadPA 42 10.5 6092 17.5 32 161 1015.0

+ pittsbPA 36 10.6 3437 8.1 59 263 991.3

+ portldOR 37 12.0 3387 3.6 21 44 894.0

+ provdcRI 42 10.1 3508 2.2 4 18 938.5

+ readngPA 41 9.6 4843 2.7 11 89 946.2

+ richmdVA 44 11.0 3768 28.6 9 48 1026.0

+ rochtrNY 32 11.1 4355 5.0 4 18 874.3

+ stlousMO 34 9.7 5160 17.2 15 68 953.6

+ sandigCA 10 12.1 3033 5.9 66 20 839.7

+ sanfrnCA 18 12.2 4253 13.7 171 86 911.7

+ sanjosCA 13 12.2 2702 3.0 32 3 790.7

+ seatleWA 35 12.2 3626 5.7 7 20 899.3

+ springMA 45 11.1 1883 3.4 4 20 904.2

+ syracuNY 38 11.4 4923 3.8 5 25 950.7

+ toledoOH 31 10.7 3249 9.5 7 25 972.5

+ uticaNY 40 10.3 1671 2.5 2 11 912.2

+ washDC 41 12.3 5308 25.9 28 102 968.8

+ wichtaKS 28 12.1 3665 7.5 2 1 823.8

+ wilmtnDE 45 11.3 3152 12.1 11 42 1004.0

+ worctrMA 45 11.1 3678 1.0 3 8 895.7

+ yorkPA 42 9.0 9699 4.8 8 49 911.8

+ youngsOH 38 10.7 3451 11.7 13 39 954.4

+ ", sep=" ")

> options(scipen=999) # suppressing scientific notation

> airpol.full <- read.table(my.datafile, header=FALSE, col.names=c("City", "Rainfall", "Education", "Popden", "Nonwhite", "NOX", "SO2", "Mortality"))

> View(airpol.full)

> airpol.full

City Rainfall Education Popden Nonwhite NOX SO2 Mortality

1 akronOH 36 11.4 3243 8.8 15 59 921.9

2 albanyNY 35 11.0 4281 3.5 10 39 997.9

3 allenPA 44 9.8 4260 0.8 6 33 962.4

4 atlantGA 47 11.1 3125 27.1 8 24 982.3

5 baltimMD 43 9.6 6441 24.4 38 206 1071.0

6 birmhmAL 53 10.2 3325 38.5 32 72 1030.0

7 bostonMA 43 12.1 4679 3.5 32 62 934.7

8 bridgeCT 45 10.6 2140 5.3 4 4 899.5

9 bufaloNY 36 10.5 6582 8.1 12 37 1002.0

10 cantonOH 36 10.7 4213 6.7 7 20 912.3

11 chatagTN 52 9.6 2302 22.2 8 27 1018.0

12 chicagIL 33 10.9 6122 16.3 63 278 1025.0

13 cinnciOH 40 10.2 4101 13.0 26 146 970.5

14 clevelOH 35 11.1 3042 14.7 21 64 986.0

15 colombOH 37 11.9 4259 13.1 9 15 958.8

16 dallasTX 35 11.8 1441 14.8 1 1 860.1

17 daytonOH 36 11.4 4029 12.4 4 16 936.2

18 denverCO 15 12.2 4824 4.7 8 28 871.8

19 detrotMI 31 10.8 4834 15.8 35 124 959.2

20 flintMI 30 10.8 3694 13.1 4 11 941.2

21 ftwortTX 31 11.4 1844 11.5 1 1 891.7

22 grndraMI 31 10.9 3226 5.1 3 10 871.3

23 grnborNC 42 10.4 2269 22.7 3 5 971.1

24 hartfdCT 43 11.5 2909 7.2 3 10 887.5

25 houstnTX 46 11.4 2647 21.0 5 1 952.5

26 indianIN 39 11.4 4412 15.6 7 33 968.7

27 kansasMO 35 12.0 3262 12.6 4 4 919.7

28 lancasPA 43 9.5 3214 2.9 7 32 844.1

29 losangCA 11 12.1 4700 7.8 319 130 861.8

30 louisvKY 30 9.9 4474 13.1 37 193 989.3

31 memphsTN 50 10.4 3497 36.7 18 34 1006.0

32 miamiFL 60 11.5 4657 13.5 1 1 861.4

33 milwauWI 30 11.1 2934 5.8 23 125 929.2

34 minnplMN 25 12.1 2095 2.0 11 26 857.6

35 nashvlTN 45 10.1 2082 21.0 14 78 961.0

36 newhvnCT 46 11.3 3327 8.8 3 8 923.2

37 neworlLA 54 9.7 3172 31.4 17 1 1113.0

38 newyrkNY 42 10.7 7462 11.3 26 108 994.6

39 philadPA 42 10.5 6092 17.5 32 161 1015.0

40 pittsbPA 36 10.6 3437 8.1 59 263 991.3

41 portldOR 37 12.0 3387 3.6 21 44 894.0

42 provdcRI 42 10.1 3508 2.2 4 18 938.5

43 readngPA 41 9.6 4843 2.7 11 89 946.2

44 richmdVA 44 11.0 3768 28.6 9 48 1026.0

45 rochtrNY 32 11.1 4355 5.0 4 18 874.3

46 stlousMO 34 9.7 5160 17.2 15 68 953.6

47 sandigCA 10 12.1 3033 5.9 66 20 839.7

48 sanfrnCA 18 12.2 4253 13.7 171 86 911.7

49 sanjosCA 13 12.2 2702 3.0 32 3 790.7

50 seatleWA 35 12.2 3626 5.7 7 20 899.3

51 springMA 45 11.1 1883 3.4 4 20 904.2

52 syracuNY 38 11.4 4923 3.8 5 25 950.7

53 toledoOH 31 10.7 3249 9.5 7 25 972.5

54 uticaNY 40 10.3 1671 2.5 2 11 912.2

55 washDC 41 12.3 5308 25.9 28 102 968.8

56 wichtaKS 28 12.1 3665 7.5 2 1 823.8

57 wilmtnDE 45 11.3 3152 12.1 11 42 1004.0

58 worctrMA 45 11.1 3678 1.0 3 8 895.7

59 yorkPA 42 9.0 9699 4.8 8 49 911.8

60 youngsOH 38 10.7 3451 11.7 13 39 954.4

> View(airpoll.full)

> is.data.frame(airpol.full)

[1] TRUE

> summary.data.frame(airpol.full)

City Rainfall Education Popden Nonwhite NOX SO2

akronOH : 1 Min. :10.00 Min. : 9.00 Min. :1441 Min. : 0.80 Min. : 1.00 Min. : 1.00

albanyNY: 1 1st Qu.:32.75 1st Qu.:10.40 1st Qu.:3104 1st Qu.: 4.95 1st Qu.: 4.00 1st Qu.: 11.00

allenPA : 1 Median :38.00 Median :11.05 Median :3567 Median :10.40 Median : 9.00 Median : 30.00

atlantGA: 1 Mean :37.37 Mean :10.97 Mean :3866 Mean :11.87 Mean : 22.65 Mean : 53.77

baltimMD: 1 3rd Qu.:43.25 3rd Qu.:11.50 3rd Qu.:4520 3rd Qu.:15.65 3rd Qu.: 23.75 3rd Qu.: 69.00

birmhmAL: 1 Max. :60.00 Max. :12.30 Max. :9699 Max. :38.50 Max. :319.00 Max. :278.00

(Other) :54

Mortality

Min. : 790.7

1st Qu.: 898.4

Median : 943.7

Mean : 940.4

3rd Qu.: 983.2

Max. :1113.0

**Then use the following code to “extract” the first 16 observations.**

city.names <- as.character(airpol.full[1:16,1])

airpol.data.sub <- airpol.full[1:16,2:8]

**Display the subset data airpol.data.sub**

> city.names <- as.character(airpol.full[1:16,1])

> airpol.data.sub <- airpol.full[1:16,2:8]

> city.names

[1] "akronOH" "albanyNY" "allenPA" "atlantGA" "baltimMD" "birmhmAL" "bostonMA" "bridgeCT" "bufaloNY"

[10] "cantonOH" "chatagTN" "chicagIL" "cinnciOH" "clevelOH" "colombOH" "dallasTX"

> airpol.data.sub

Rainfall Education Popden Nonwhite NOX SO2 Mortality

1 36 11.4 3243 8.8 15 59 921.9

2 35 11.0 4281 3.5 10 39 997.9

3 44 9.8 4260 0.8 6 33 962.4

4 47 11.1 3125 27.1 8 24 982.3

5 43 9.6 6441 24.4 38 206 1071.0

6 53 10.2 3325 38.5 32 72 1030.0

7 43 12.1 4679 3.5 32 62 934.7

8 45 10.6 2140 5.3 4 4 899.5

9 36 10.5 6582 8.1 12 37 1002.0

10 36 10.7 4213 6.7 7 20 912.3

11 52 9.6 2302 22.2 8 27 1018.0

12 33 10.9 6122 16.3 63 278 1025.0

13 40 10.2 4101 13.0 26 146 970.5

14 35 11.1 3042 14.7 21 64 986.0

15 37 11.9 4259 13.1 9 15 958.8

16 35 11.8 1441 14.8 1 1 860.1

**Part 2:** Use R to perform the following analysis on the subset data airpol.data.sub. Make sure you include clear headings, command lines, and relevant output/results.

1. Calculate the sample covariance matrix and the sample correlation matrix. Identify which pairs of variables seem to be strongly associated, and describe the nature (strength and direction) of the relationship between these variable pairs.

# created sub set called air.data.cor for sample correlation matrix, rounded the matric output to 2 zero after decimal point and using the correlation function within R “cor ()”.

## red means high correlation and yellow means medium correlation, based on the output the SO2 ↔NOX had a very close linear relationship when one goes up or down the other will have similar affect, while SO2 ↔ Popden also have linear relationship not quite as strong as SO2 and NOX but strong enough to assume a possible linear relationship between the two variables.

> airpol.data.cor <- round(cor(airpol.data.sub), digits = 2)

> airpol.data.cor

Rainfall Education Popden Nonwhite NOX SO2 Mortality

Rainfall 1.00 -0.50 -0.29 0.54 -0.08 -0.17 0.29

Education -0.50 1.00 -0.19 -0.28 -0.09 -0.27 -0.58

Popden -0.29 -0.19 1.00 -0.11 0.58 0.62 0.58

Nonwhite 0.54 -0.28 -0.11 1.00 0.31 0.25 0.53

NOX -0.08 -0.09 0.58 0.31 1.00 0.92 0.57

SO2 -0.17 -0.27 0.62 0.25 0.92 1.00 0.59

Mortality 0.29 -0.58 0.58 0.53 0.57 0.59 1.00

# SO2 ↔NOX covariance is positive (1169.95)

# SO2 ↔ Popden covariance is positive (71437.88)

## this means the scores tend to covary in a positive way. As scores on SO2 go up, scores on NOX and Popden also tend to go up; and vice versa.

> airpol.data.cov <- round(cov(airpol.data.sub), digits = 2)

> airpol.data.cov

Rainfall Education Popden Nonwhite NOX SO2 Mortality

Rainfall 39.72 -2.46 -2766.77 34.61 -8.30 -84.69 101.49

Education -2.46 0.61 -224.44 -2.25 -1.14 -16.09 -24.96

Popden -2766.77 -224.44 2229957.93 -1665.31 14294.73 71437.88 47577.01

Nonwhite 34.61 -2.25 -1665.31 102.69 51.96 198.87 294.06

NOX -8.30 -1.14 14294.73 51.96 268.60 1169.95 513.19

SO2 -84.69 -16.09 71437.88 198.87 1169.95 5981.26 2502.85

Mortality 101.49 -24.96 47577.01 294.06 513.19 2502.85 3030.05

1. Calculate the distance matrix for these observations (after scaling the variables by dividing each variable by its standard deviation). Describe some of the most similar pairs of cities and some of the most different pairs of cities, giving evidence from the distance matrix.

> std <- sapply(airpol.data.sub, sd)

> std

Rainfall Education Popden Nonwhite NOX SO2 Mortality

6.3021160 0.7841928 1493.3043673 10.1338377 16.3890207 77.3386223 55.0459142

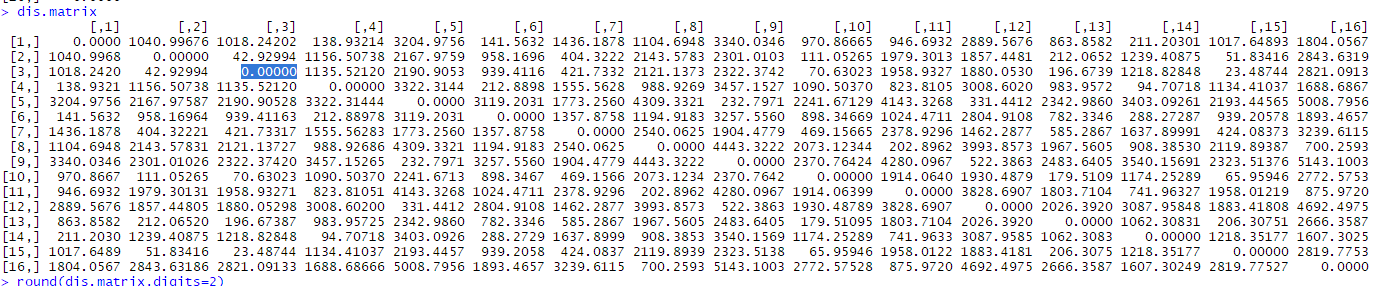
> airpol.data.std <- sweep(airpol.data.sub, 2, std, FUN = "/")

> dis <- dist(airpol.data.sub)

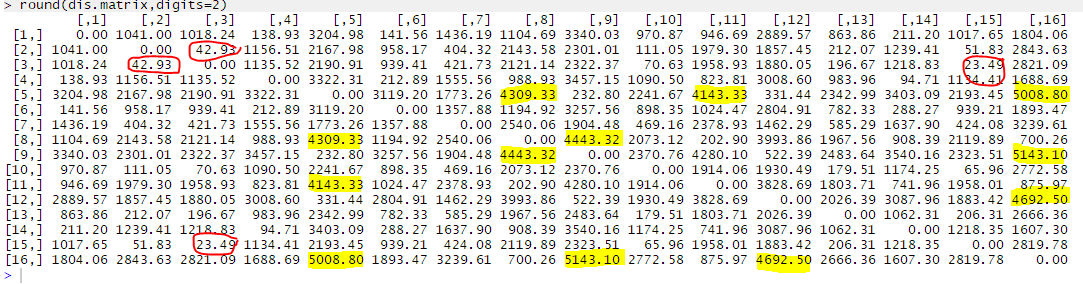
> airpol.data.std

> dis.matrix<-dist2full(dis)

> dis.matrix



> round(dis.matrix,digits=2)



1. Display a plot that will help assess whether this data set comes from a multivariate normal distribution. What is your conclusion based on the plot?

#all plots within this dataset follow approximately linear model and look to follow the rules for normality, we found a few outliers that will need to be considered for the following dataset: Popden, Nonwhite, NOX and SO2.

##Overall the data looks to follow the normality guidelines

> qqnorm(airpol.data.sub[,1], ylab="Ordered Observations", main = "Normal Q-Q Plot, Rainfall")

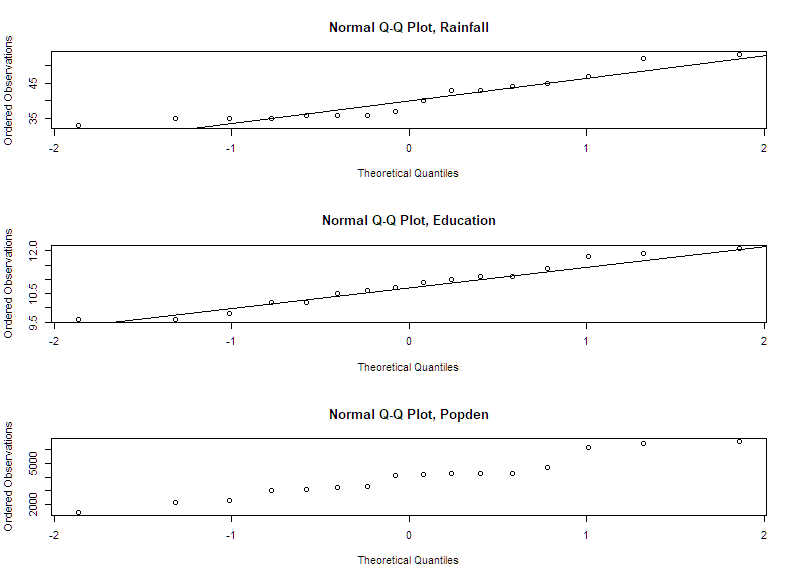
> qqline(airpol.data.sub[,1])

> qqnorm(airpol.data.sub[,2], ylab="Ordered Observations", main = "Normal Q-Q Plot, Education")

> qqline(airpol.data.sub[,2])

> qqnorm(airpol.data.sub[,3], ylab="Ordered Observations", main = "Normal Q-Q Plot, Popden")

> qqline(airpol.data.sub[,2])



> qqnorm(airpol.data.sub[,4], ylab="Ordered Observations", main = "Normal Q-Q Plot, Nonwhite")

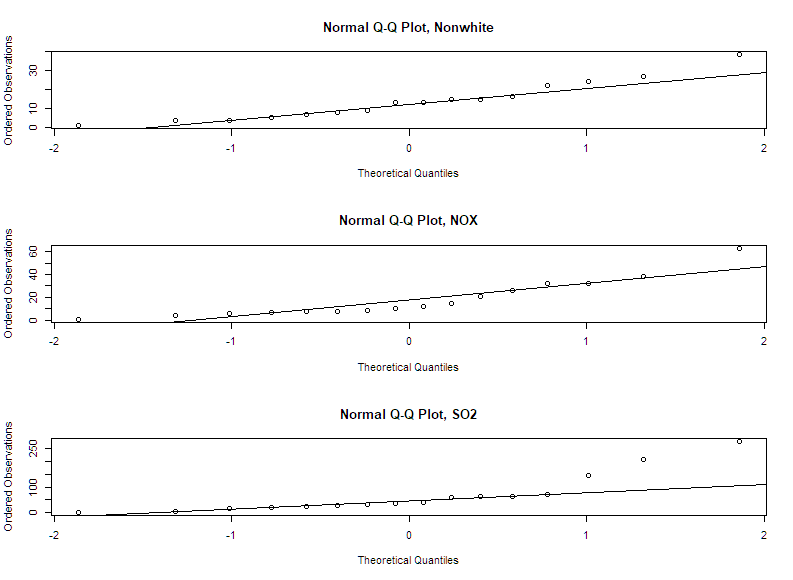
> qqline(airpol.data.sub[,4])

> qqnorm(airpol.data.sub[,5], ylab="Ordered Observations", main = "Normal Q-Q Plot, NOX")

> qqline(airpol.data.sub[,5])

> qqnorm(airpol.data.sub[,6], ylab="Ordered Observations", main = "Normal Q-Q Plot, SO2")

> qqline(airpol.data.sub[,6])



> qqnorm(airpol.data.sub[,7], ylab="Ordered Observations", main = "Normal Q-Q Plot, Mortality")

> qqline(airpol.data.sub[,7])

